A Variable Selection Method for Analyzing Supersaturated Designs

Christos Koukouvinos, K. Mylona, Anna Skountzou

A Variable Selection Method for Analyzing Supersaturated Designs

Journal:	*Communications in Statistics - Simulation and Computation*
Manuscript ID:	LSSP-2010-0077.R1
Manuscript Type:	Original Paper
Date Submitted by the Author:	28-Oct-2010
Complete List of Authors:	Koukouvinos, Christos; National Technical University of Athens Mylona, K.; Universiteit Antwerpen Skountzou, Anna; National Technical University of Athens
Keywords:	Supersaturated designs, Variable selection, Best subset method, Information criteria
Abstract:	Supersaturated designs are a large class of factorial designs which can be used for screening out the important factors from a large set of potentially active variables. The huge advantage of these designs is that they reduce the experimental cost drastically, but their critical disadvantage is the confounding involved in the statistical analysis. In this paper, we propose a method for analyzing data using several types of supersaturated designs. Modifications of widely used information criteria are given and applied to the variable selection procedure for the identification of the active factors.
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
LSSP-2010-0077.zip	

A Variable Selection Method for Analyzing Supersaturated Designs

C. Koukouvinos[1], K. Mylona[2] and A. Skountzou[1]
[1]Department of Mathematics, National Technical University of Athens,
Zografou 15773, Athens, Greece.
[2]Faculty of Applied Economics, Universiteit Antwerpen,
City Campus - Prinsstraat 13, 2000 Antwerpen, Belgium.

Abstract

Supersaturated designs are a large class of factorial designs which can be used for screening out the important factors from a large set of potentially active variables. The huge advantage of these designs is that they reduce the experimental cost drastically, but their critical disadvantage is the confounding involved in the statistical analysis. In this paper, we propose a method for analyzing data using several types of supersaturated designs. Modifications of widely used information criteria are given and applied to the variable selection procedure for the identification of the active factors. The effectiveness of the proposed method is depicted via simulated experiments and comparisons.

Keywords and phrases: Supersaturated designs, Variable selection, Best subset method, Information criteria.
AMS Subject Classification: 62-07, 62K15.

1 Introduction

Supersaturated designs (SSDs) are used in the initial stage of an industrial or scientific experiment for identifying the active effects, and they are very useful when there is a large number of factors under testing while only a limited number of experimental runs is available. The analysis of supersaturated designs relies on the assumption of effect sparsity (Box and Meyer ([4])), i.e. only few of the experimental factors have significant influence on the response. SSDs can be generally described as designs with p factors and n runs where $n \leq p$. The idea of SSDs was initiated by Satterthwaite ([25]). Even though the construction methods for SSDs have been widely studied (Lin ([14], [15]), Nguyen ([20]), Tang and Wu ([27]), Liu and Zhang ([17]), Liu and Dean ([16]),

Ryan and Bulutoglu ([24]), Nguyen and Cheng ([21])), the analysis of SSDs is yet in an early research stage.

Westfall et al. ([29]) proposed an error control skill in forward selection. Abraham et al. ([1]) pointed out that the analysis of SSDs can give uncertain results independently of the method used. They applied stepwise and all-models methods to identify the active factors. A two-stage Bayesian model selection strategy for SSDs, able to keep all possible models under consideration while providing a level of robustness similar to Bayesian analysis incorporating noninformative priors, was proposed by Beattie et al. ([3]). Li and Lin ([12]) introduced a variable selection approach to identify the active effects in SSDs via nonconvex penalized least squares. To find the solution of the penalized least squares an iterative ridge regression is employed. This method is empirically compared to the Bayesian variable selection (Li and Lin ([13])). The forward selection method has been suggested and performed by Lin ([14]), where the data were based on a half fraction of Plackett-Burman design. Wang ([28]) applied the Lin's analysis on the other half fraction of Plackett-Burman design and observed that four of the five active factors found in one half fraction were not found in the other. However, when effect sparsity holds, Type I errors can easily occur. In a recent paper, Holcomb et al. ([8]) proposed the bootstrap method with contrasts and the contrast variance method to analyze data using a broad range of supersaturated designs. The objective of the contrast variance method they proposed is to reduce Type II error rates. Hamada and Wu ([6])) recommended an iterative guided stepwise regression strategy for analyzing the data from these designs that allows entertainment of interactions. However, their strategy provides a restricted search in a rather large model space. Chipman et al. ([5]) proposed a Bayesian variable selection approach for analyzing experiments with complex aliasing and Lu and Wu ([18]) proposed an analysis method based on the idea of staged dimensionality reduction. Phoa et al. ([22]) studied a variable selection method via the Dantzig selector for analyzing supersaturated designs.

The paper is organized as follows. In Section 2, we present the proposed method and implement its steps using a best subset variable selection approach. Also, we introduce two modifications of the widely used information criteria; Akaike Information Criterion (AIC, Akaike ([2])) and Bayesian Information Criterion (BIC, Schwarz ([26])). Simulation results are listed and discussed in Section 3. Some conclusion comments are summarized in the last section.

2 Proposed Method

Let us consider the model for a screening experiment

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y}$ is the $(n \times 1)$ response vector, $\mathbf{X}$ is the $n \times p$ design matrix, $\boldsymbol{\beta}$ is the $(p \times 1)$ vector of the unknown coefficients, and $\boldsymbol{\varepsilon}$ is the $(n \times 1)$ random error vector with $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ and $\varepsilon_i \sim N(0, \sigma^2)$ for all $i = 1, 2, \ldots, n$. The aim of factor screening is to identify those factors which have non-zero effects. The experimenter can assume that there are only up to m active factors from the total set of p factors involved. Usually, the process is more robust when m is at most half the number of runs (effect sparsity ([4])).

In the present framework we propose an analysis method for testing the factorial effects in a supersaturated design with a block structure. Our method is based on the procedure of subset selection used with certain criteria which combine statistical measures with penalties for increasing the number of predictors in the model. Reviews on best subset variable selection method can be found in Hocking ([7]), Miller ([4]) and Rao and Wu ([23]). Basically, the criteria used in the best subset variable selection procedure are classified into four categories: (1) Prediction criteria; (2) Information-based criteria; (3) Data-reuse and Data-driven procedures; (4) Bayesian variable selection.

We use here only the information-based criteria which are related to likelihood or divergence measures. Each of these criteria can be considered as an approximately unbiased estimator of the expected overall discrepancy, a nonnegative quantity which measures the distance between the underlying model and the fitted approximated (or candidate) model. These criteria usually consist of two terms where the first represents a biased estimator of the expected overall discrepancy and the second is the appropriate correction (or penalty) term that makes the estimator asymptotically unbiased. The most popular criteria of this class include the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). AIC was proposed by Akaike ([2]), and it selects the model that minimizes AIC $= -2l + 2p$, where l is the loglikelihood of the model and p is the number of parameters in the model. BIC was proposed by Schwarz ([26]) and has a similar form to AIC except that the log-likelihood is penalized by $plog(n)$ instead of 2p, selecting the model that minimizes BIC $= -2l + p\log(n)$, where n is the number of observations.

Phoa et al. ([22]) suggested a modified AIC criterion for model comparison, which is equivalent to mAIC $= -2l + 2p^2$. They posed a quadratic penalty of model complexity p than the linear penalty in AIC; thus, mAIC chooses more parsimonious model than AIC. Inspired from this work, we propose two more modifications of the above criteria; the modified AIC and the modified BIC which are given from the relationships modAIC $= \frac{-2l}{p} + \frac{2p^{3/2}}{\log n^{3/2}}$ and modBIC $= \frac{-2l}{p} + \frac{p \log n^2}{\log p}$, respectively. In our criteria, the conventional AIC and BIC are standardized by dividing the number of active factors and include it in the first term. Thus, the effect of the number of factors is minimized. Also, a penalized multiplier including the model complexity and the number of observations is added in both modified criteria. Several trials and simulated experiments showed

that these multipliers give the best results compared to other multipliers and the original versions of the criteria. The efficiency of these modifications is established via a comparative simulation study which is presented in Section 3.

Our method is applicable to the s-block two-level $E(s^2)$-optimal supersaturated designs with n runs and $p = s(n-1)$ factors. Such designs can be provided by the construction methods of Tang and Wu ([27]) and Koukouvinos et al. ([10],[11]).

According to the Tang and Wu method ([27]), one can construct a supersaturated design $\mathbf{X}$ with n rows and $s(n-1)$ columns by juxtaposing s equivalent Hadamard matrices $\mathbf{H_1}, \mathbf{H_2}, \ldots, \mathbf{H_s}$ of order n, with the requirement that none of the $s(n-1)$ columns be fully aliased with another (two columns $\mathbf{c_i}$ and $\mathbf{c_j}$ are said to be fully aliased if $\mathbf{c_i} = \pm\mathbf{c_j}$). Indeed, these designs can be represented by the design matrix $\mathbf{X} = [\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_s}]$, where $\mathbf{X_i}$ is a balanced $n \times (n-1)$ matrix with elements ± 1, $i = 1, 2, \ldots, s$. The criterion of minimization of $r_{max} = max_{i<j} |s_{ij}/n|$, where $s_{ij}/n = r_{ij}$ is the correlation of two columns $\mathbf{c_i}$, $\mathbf{c_j}$, is taken into account in the construction of these SSDs. Naturally, the matrix $\mathbf{X_i}$, $i = 1, 2, \ldots, s$ has its columns pairwise orthogonal. An important feature of these designs is that any two columns which have non zero correlation appear in different blocks. So, the correlations of all pairs of columns within each block of the design matrix is equal to zero. In this work, the proposed method takes advantage of this feature, having as goal to overcome problems arising if a column of the design is correlated with another column and both correspond to active factors.

Koukouvinos et al. ([10]) construct $E(s^2)$-optimal and minimax-optimal, two level cyclic structured supersaturated designs through a metaheuristic approach guided via multi-objective simulated annealing. Although the blocks constituting these designs have not necessarily an orthogonal structure, these SSDs appear optimal with respect to the criterion of minimization of r_{max}. In this way the value of the appeared correlation between the columns is as small as possible. Also, Koukouvinos et al. ([11]) propose a hybrid simulated annealing genetic algorithm for generating $E(s^2)$-optimal cyclic structured supersaturated designs.

The procedure we suggest is briefly described by the following steps.

1. For $i = 1, 2, \ldots, s$ apply the best subset variable selection method, combined with one of the information criteria described previously, using as design matrix the $n \times (n-1)$ matrix $\mathbf{X_i}$, and the $n \times 1$ response vector $\mathbf{y}$.

2. Define S_i be the set of all active variables derived when $\mathbf{y}$ regress with $\mathbf{X_i}$, $i = 1, 2, \ldots, s$.

3. Let $S = \bigcup_{i=1}^{s} S_i$ is the union of all factors derived in each block of the design and consider the set of screened active variables to be S.

This method is applied to both type of designs as well as to SSDs without a block structure ([20]) with very good results presented in the next section.

3 Simulation Study

In this section a comparative simulation study is provided in order to test the performance of the proposed method. The SCAD method ([12],([9])) is also included in the simulation experiments, so as the best subset variable selection method is compared to an existing analysis method of supersaturated designs. Since the proposed method is run into blocks, SCAD is also run into blocks as it is proposed in [9]. The necessary parameters for the SCAD method are selected according to the suggestion in the original paper they appear. Specifically, we choose $a = 3.7$ and a generalized cross-validation is used to estimate the thresholding parameter λ.

1000 data sets are generated from the linear model (1) with randomly selected coefficients and an $\varepsilon_i \sim N(0,1)$ for all $i = 1, 2, \ldots, n$ random error is added to each corresponding observation y_i, while only main effects models are considered. The true active variables are also selected randomly from the set of $1, \ldots, p$ potentially active factors, with respect to the already determined number of active variables in each block of the design matrix. Following the principle of effect sparsity, all possible numbers of active factors are chosen. Thus, $0, 1, 2, 3, \ldots, \frac{n}{2}$, active factors are tested for each of the six $n \times p$ design matrices we study. We run all the possible combinations of $0, 1, 2, 3, \ldots, \frac{n}{2}$ active factors per s blocks in order to examine all the possible cases. We have observed symmetry between the numbers of active factors in the several blocks; hence, we limited the results listed only for the cases where $t_i < t_j$, $i \neq j$, $i, j = 1, ..., s$ and t_i denote the number of active factors in the i-th block. The coefficients of the non-active variables, in the true model, are generated from $N(0, 0.05)$ and the absolute values of the active factors are randomly selected in the range $[0.7, 2.0]$, considering both negative and positive signs assigned to the coefficients of the active factors. In total, for obtaining Tables I-V the number of the used linear models is equal to the number of the considered cases appeared in the corresponding tables.

In these simulation experiments our aim is to control two types of error rates; the cost of declaring an inert effect to be active (Type I error), and the cost of declaring an active effect to be inactive (Type II error).

Various types of supersaturated designs are used in the simulations as examples of the performance of the method when it is applied in designs with different structure. In Tables I-V are presented the results obtained by one supersaturated design constructed by the method presented in Tang and Wu ([27]), two supersaturated designs constructed by Koukouvinos et al. ([10]), one supersaturated design constructed by Koukouvinos et al. ([11]) and one supersaturated design

constructed by Nguyen ([20]), respectively. In the first columns of each table we present the number of true active factors t_i, $i = 1, \ldots, s$, per each of the s blocks used in the simulated models. In the next columns we present the Type I and Type II error rates occurred by the application of the best subset variable selection and the block SCAD ([9]) method to the simulated models. The procedure of best subset variable selection terminates using the five information-based criteria described in Section 2.

- **A two-block orthogonal SSD**

The first idea was to apply the proposed method to a two-block orthogonal design in order to take advantage of this special structure. We choose the $E(s^2)$-optimal SSD for $p = 22$ factors in $n = 12$ runs with $r_{max} = 0.67$ constructed in Tang and Wu ([27]) for this purpose and the results are presented in the following table.

t		AIC		BIC		mAIC		modAIC		modBIC		SCAD	
t_1	t_2	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
0	0	0.1639	-	0.1652	-	0.0781	-	0.0792	-	0.0835	-	0.1324	-
0	1	0.2279	0.0980	0.2310	0.0960	0.0615	0.2270	0.1224	0.1800	0.1345	0.1740	0.1604	0.1290
0	2	0.3033	0.0630	0.3070	0.0615	0.0696	0.3535	0.1620	0.1215	0.1875	0.1060	0.1829	0.1175
1	1	0.3245	0.0660	0.3274	0.0635	0.0559	0.2850	0.1698	0.1350	0.1983	0.1190	0.1962	0.1540
0	3	0.3583	0.0950	0.3613	0.0933	0.0835	0.4940	0.1819	0.1907	0.2081	0.1590	0.2132	0.1867
1	2	0.3890	0.0307	0.3911	0.0293	0.0549	0.2150	0.1798	0.0773	0.2164	0.0637	0.2071	0.1093
0	4	0.3931	0.0602	0.3947	0.0587	0.1057	0.5383	0.1998	0.1838	0.2189	0.1358	0.2277	0.1417
1	3	0.3801	0.0305	0.3815	0.0302	0.0670	0.3342	0.1785	0.1038	0.2011	0.0800	0.2147	0.1525
2	2	0.4095	0.0073	0.4116	0.0073	0.0436	0.2190	0.1802	0.0382	0.2112	0.0267	0.2165	0.0685
0	5	0.4725	0.0564	0.4747	0.0562	0.1160	0.6096	0.2262	0.2222	0.2678	0.1668	0.2675	0.2182
1	4	0.4836	0.0874	0.4849	0.0870	0.0686	0.4396	0.2085	0.1976	0.2594	0.1682	0.2126	0.2412
2	3	0.4999	0.1044	0.5020	0.1036	0.1181	0.6462	0.2515	0.2648	0.3025	0.3142	0.3245	0.4296
0	6	0.4538	0.0467	0.4549	0.0450	0.1251	0.6695	0.2432	0.2922	0.2592	0.2087	0.2527	0.3732
1	5	0.5336	0.0712	0.5339	0.0712	0.0810	0.5508	0.2123	0.2392	0.2607	0.1950	0.1773	0.3293
2	4	0.5208	0.1060	0.5216	0.1053	0.0750	0.5372	0.2236	0.3017	0.2738	0.2517	0.2326	0.3812
3	3	0.4079	0.0437	0.4095	0.0433	0.0504	0.4680	0.1840	0.1603	0.2243	0.1232	0.2330	0.1527

Table I: Methods performance for random model coefficients using 1000 simulations in Tang-Wu design for $p = 22$ factors and $n = 12$ runs, considering up to 6 active factors per block (11 factors in total per block).

From Table I, it is shown that the best subset variable selection methods using AIC and BIC behave similarly overestimating the non-active variables (high Type I error rates), while the application of mAIC results to extreme values of Type II. The criteria modAIC and modBIC improve the results of the corresponding original versions and compete the results obtained from block SCAD.

- **A two-block SSD**

Next, we apply the proposed method in a two-block supersaturated design provided in [10]. This design is the $E(s^2)$-optimal SSD for $p = 18$ factors in $n = 10$ runs with $r_{max} = 0.600$. The results of the simulations using the above design are listed in Table II.

t		AIC		BIC		mAIC		modAIC		modBIC		SCAD	
t_1	t_2	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
0	0	0.3668	-	0.4317	-	0.0912	-	0.0997	-	0.1479	-	0.1271	-
0	1	0.4706	0.1370	0.5211	0.1300	0.0741	0.2880	0.1360	0.2090	0.2134	0.2130	0.1259	0.2140
0	2	0.5691	0.1225	0.6025	0.1210	0.0839	0.4405	0.1916	0.2155	0.3064	0.1800	0.1124	0.2185
1	1	0.5671	0.3385	0.6163	0.3350	0.0861	0.5875	0.2028	0.4730	0.2938	0.4250	0.1509	0.5215
0	3	0.4771	0.1540	0.5145	0.1483	0.0897	0.5990	0.2003	0.3097	0.2726	0.2377	0.1835	0.3427
1	2	0.5231	0.1863	0.5544	0.1773	0.0917	0.5493	0.2246	0.3583	0.3114	0.2820	0.1905	0.4243
0	4	0.5876	0.1330	0.6120	0.1270	0.1139	0.6005	0.2457	0.3113	0.3620	0.2115	0.1487	0.3945
1	3	0.5692	0.1403	0.5999	0.1320	0.0821	0.4808	0.2390	0.2858	0.3405	0.2218	0.1746	0.4060
2	2	0.6841	0.1452	0.7059	0.1260	0.1024	0.5517	0.2886	0.4035	0.4314	0.3225	0.1269	0.5357
0	5	0.6876	0.1090	0.6978	0.1062	0.1492	0.6284	0.2866	0.3402	0.4577	0.2194	0.1045	0.4628
1	4	0.7327	0.1784	0.7492	0.1698	0.1452	0.6238	0.3385	0.4838	0.4782	0.3914	0.1336	0.5344
2	3	0.6419	0.3150	0.6636	0.2898	0.1597	0.7452	0.3432	0.5186	0.4375	0.5334	0.2445	0.6676

Table II: Methods performance for random model coefficients using 1000 simulations in Koukouvinos et al. design for $p = 18$ factors and $n = 10$ runs, considering up to 5 active factors per block (9 factors in total per block).

The previous observations stand for the results of Table II too; when best subset terminates using AIC and BIC, the values of Type I are very high, while the use of mAIC results to extreme Type II error rates. In contrary, the modified versions modAIC and modBIC behave better than AIC and BIC, since they reduce the values of Type I and also, give lower Type II rates than SCAD.

- **A larger SSD**

In cases that the number of experimental factors is large, best subset approaches could not be applied because of their computer burden. So, it is of interest to test if a further division within the blocks could be an alternative approach in such cases. We use the supersaturated design for $p = 66$ factors and $n = 12$ runs ([10]) and we divide each one of the two original blocks into three blocks; so, we have six blocks in total. For simplicity, we keep the same notation for the number of active factors as in previous tables, but note that we now refer in a subdivision of the two-block design. Active factors are assigned randomly in each of the six sub-blocks.

t		AIC		BIC		mAIC		modAIC		modBIC		SCAD	
t_1	t_2	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
0	0	0.2118	-	0.2301	-	0.0497	-	0.0838	-	0.1360	-	0.0861	-
0	1	0.3627	0.1820	0.3788	0.1820	0.0814	0.2140	0.1236	0.1840	0.2363	0.1800	0.1381	0.1830
0	2	0.4252	0.2250	0.4394	0.2285	0.1068	0.4565	0.1249	0.2595	0.2723	0.2350	0.1676	0.2880
1	1	0.4698	0.0860	0.4790	0.0900	0.1100	0.1575	0.1496	0.1270	0.3057	0.1115	0.1573	0.1095
0	3	0.5090	0.2220	0.5171	0.2220	0.1327	0.4367	0.1750	0.3360	0.3288	0.3090	0.1839	0.3807
1	2	0.4794	0.2743	0.4888	0.2687	0.1305	0.4853	0.1690	0.3897	0.3228	0.3570	0.1903	0.4350
0	4	0.5622	0.0747	0.5707	0.0752	0.1395	0.2288	0.1794	0.1358	0.3377	0.1155	0.1899	0.2208
1	3	0.5842	0.1318	0.5896	0.1345	0.1462	0.3320	0.1868	0.1683	0.3548	0.1452	0.1880	0.2402
2	2	0.5831	0.0210	0.5874	0.0205	0.1352	0.1817	0.1622	0.1678	0.3603	0.0532	0.1781	0.1920
0	5	0.6351	0.1642	0.6402	0.1618	0.1504	0.4744	0.2140	0.3032	0.3703	0.2662	0.1991	0.4236
1	4	0.6881	0.1066	0.6925	0.1054	0.1393	0.3046	0.1736	0.2208	0.3893	0.2008	0.1857	0.2656
2	3	0.6748	0.0740	0.6792	0.0728	0.1413	0.3278	0.2283	0.2514	0.3931	0.2110	0.2118	0.3432
0	6	0.6545	0.1822	0.6577	0.1793	0.1570	0.5843	0.2034	0.3870	0.4071	0.3350	0.2033	0.4458
1	5	0.6989	0.0952	0.7020	0.0962	0.1391	0.3933	0.2263	0.1690	0.3983	0.1387	0.2006	0.3938
2	4	0.7018	0.0820	0.7047	0.0812	0.1434	0.4357	0.2096	0.3227	0.4092	0.2388	0.2109	0.3810
3	3	0.6438	0.1197	0.6480	0.1190	0.1493	0.4912	0.2216	0.2955	0.4001	0.2263	0.2311	0.3705

Table III: Methods performance for random model coefficients using 1000 simulations in Koukouvinos et al. design for $p = 66$ factors and $n = 12$ runs, considering up to 6 active factors per block (11 factors in total in each of the six blocks). Each one of the two original blocks is sub-divided in three blocks.

Let us note that the further split of the design in six blocks is less time consuming than the division in two blocks. The modified versions *mod*AIC and *mod*BIC improve the results of the original versions. *mod*BIC yields lower Type II error rates than SCAD, but higher values of Type I. On the other hand, *mod*AIC outperforms SCAD run into blocks. The above results show that when the number of experimental factors enlarges a further division into sub-blocks could be an alternative approach of the proposed best subset method using *mod*AIC.

- **A three-block SSD**

We also use the $E(s^2)$-optimal SSD for $p = 21$ factors in $n = 8$ runs with $r_{max} = 0.500$ ([11]) divided in three blocks in this simulation study. The obtained results are shown in Table IV, where t_1, t_2, t_3 denote the number of active factors in the first, second and third block respectively. Again *mod*AIC and *mod*BIC perform better than AIC and BIC giving lower Type I rates and preserving Type II rates in satisfactory levels. It is shown also, that *mod*AIC gives better results than SCAD run into three blocks.

t			AIC		BIC		mAIC		*mod*AIC		*mod*BIC		SCAD	
t_1	t_2	t_3	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
0	0	0	0.3672	-	0.4120	-	0.0820	-	0.0963	-	0.2160	-	0.1053	-
1	0	0	0.4501	0.4290	0.4894	0.3950	0.0916	0.6040	0.1943	0.4580	0.2882	0.3640	0.2182	0.5020
2	0	0	0.6033	0.2470	0.6310	0.2370	0.1345	0.5570	0.2104	0.3450	0.3962	0.1725	0.2588	0.4575
1	1	0	0.5761	0.2955	0.6127	0.2720	0.1162	0.5185	0.2017	0.4995	0.3808	0.2880	0.2593	0.5270
3	0	0	0.5723	0.3157	0.5975	0.2927	0.1553	0.6587	0.2183	0.4827	0.3846	0.2603	0.2981	0.5970
2	1	0	0.6593	0.3243	0.6792	0.2963	0.1676	0.6433	0.2954	0.6203	0.4681	0.3447	0.2828	0.6143
1	1	1	0.6889	0.3237	0.7139	0.3077	0.1196	0.4070	0.2958	0.4810	0.4851	0.3083	0.3298	0.4090
4	0	0	0.6421	0.3115	0.6615	0.2953	0.1886	0.7083	0.2728	0.5058	0.4378	0.2643	0.3305	0.6138
3	1	0	0.6121	0.2432	0.6341	0.2350	0.1509	0.5433	0.2505	0.3927	0.4136	0.2013	0.2954	0.4970
2	2	0	0.6209	0.0850	0.6491	0.0818	0.1476	0.2630	0.2016	0.1535	0.4049	0.1108	0.2395	0.3200
2	1	1	0.6521	0.0153	0.6760	0.0115	0.0851	0.1820	0.2249	0.1143	0.4302	0.1362	0.2276	0.2330

Table IV: Methods performance for random model coefficients using 1000 simulations in Koukouvinos et al. design for $p = 21$ factors and $n = 8$ runs, considering up to 4 active factors per block (7 factors in total per block).

- **A design without block structure**

Moreover, supersaturated designs without a block structure are used in the simulations in order to illustrate the method's conduct when it is applied to a broader class of such designs. In the following table, the results obtained from Nguyen's ([20]) ave(s^2)-optimal SSD for $p = 14$ factors in $n = 8$ runs with $r_{max} = 0.500$ are listed. The following results show that the proposed method performs well when it terminates using *mod*AIC and *mod*BIC. Moreover, best subset using *mod*AIC gives better results than block SCAD.

t		AIC		BIC		mAIC		modAIC		modBIC		SCAD	
t_L	t_R	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
0	0	0.2436	-	0.2478	-	0.1237	-	0.1260	-	0.1483	-	0.1550	-
0	1	0.3229	0.1240	0.3321	0.1210	0.0971	0.2740	0.1728	0.2190	0.2235	0.1980	0.1694	0.2220
0	2	0.3983	0.1000	0.4067	0.0945	0.0962	0.4750	0.2249	0.1950	0.3022	0.1515	0.2339	0.2435
1	1	0.3908	0.1685	0.3998	0.1580	0.1274	0.5710	0.2302	0.3405	0.2886	0.3675	0.2564	0.3750
0	3	0.4511	0.1567	0.4590	0.1487	0.1076	0.6183	0.2576	0.2950	0.3503	0.2170	0.2825	0.3900
1	2	0.4631	0.2107	0.4735	0.2040	0.0740	0.4903	0.2545	0.3220	0.3590	0.2640	0.2836	0.3983
0	4	0.5264	0.1245	0.5322	0.1210	0.1308	0.6775	0.3038	0.3040	0.4211	0.1918	0.2951	0.4895
1	3	0.4824	0.1457	0.4897	0.1442	0.0657	0.4140	0.2495	0.2275	0.3701	0.1840	0.2643	0.2920
2	2	0.5137	0.2205	0.5220	0.2147	0.1179	0.6228	0.3064	0.4022	0.4144	0.3020	0.3151	0.4150

Table VI: Methods performance for random model coefficients using 1000 simulations in Nguyen design for $p = 14$ factors and $n = 8$ runs, considering up to 4 active factors per block (7 factors in total per block).

Some general conclusion remarks are discussed in the next section.

4 Conclusion

In this paper, we have introduced a method for analyzing data in supersaturated designs by dividing them into blocks. We approach the problem of variable selection combining several information criteria, with the block structure of the SSDs given by Tang and Wu ([27]) and Koukouvinos et al. ([10], [11]). The proposed method enables us to make the best subset variable selection procedure applicable even in high order designs. However, when there is a large number of experimental runs, best subset approaches cannot be applied because of their time complexity. In these cases a further division of the large designs into more blocks allows the application of best subset methods. Empirical performance of our method based on simulations is tested and compared to SCAD run into blocks ([9]). Simulation results show that the proposed method gives good results and can compete existing analysis methods when it is applied to supersaturated designs with a block orthogonal structure. Moreover, it outperforms when it is applied to supersaturated designs with a block but not orthogonal structure or even to designs without any block structure by simply dividing them into blocks.

As with any method for analyzing supersaturated designs, high risk is expected for this method as well. Both Type I and Type II error rates are important and should be kept as low as possible. A method with a low Type I error rate has the ability to exclude unnecessary factors, so it can be helpful in reducing the cost of additional experiments based on the selected factors. On the other hand, there are experiments, where the exclusion of one or more active factors and the weakness of the method to identify them could have dramatic results. In these cases, the cost of declaring an active effect to be inactive would be substantial and the application of any of the other methods with lower Type II error rates is suggested. Best subset approach using AIC and BIC gives very high Type I and low Type II error rates. The proposed method using mAIC gives the lowest Type I error rates but extreme values of Type II. The combination of our method with the proposed in

this paper modified information criteria achieves lower error rates than the conventional versions, and especially *mod*AIC gives better results than block SCAD in most of the cases.

Acknowledgements: The authors would like to thank an Associate Editor and the anonymous referees for their useful comments and suggestions.

References

[1] Abraham, B., Chipman, H., and K. Vijayan (1999). "Some Risks in the Construction and Analysis of Supersaturated Designs". *Technometrics*, Vol. 41, pp. 135-141.

[2] H. Akaike (1969). "Fitting Autoregressive Models for Prediction". *Ann. Inst. Statist. Math.*, Vol. 21, pp. 243-247.

[3] Beattie, S. D., Fong, D. K. F., and D. K. J. Lin (2002). "A Two-Stage Bayesian Model Selection Strategy for Supersaturated Designs". *Technometrics*, Vol. 44, pp. 55-63.

[4] Box, G. E. P. and R. D. Meyer (1986). "An Analysis for Unreplicated Fractional Factorials". *Technometrics*, Vol. 28, pp. 11-18.

[5] Chipman, H., Hamada, M., and C. F. J. Wu (1997). "A Bayesian Variable Selection Approach for Analyzing Designed Experiments with Complex Aliasing". *Technometrics*, Vol. 39, pp. 372-381.

[6] Hamada, M., and C. F. J. Wu (1992). "Analysis of Designed Experiments with Complex Aliasing". *Journal of Quality Technology*, Vol. 24, pp. 130-137.

[7] R. Hocking (1976). "The Analysis and Selection of Variables in Linear Regression". *Biometrics*, Vol. 32, pp. 1-49.

[8] Holcomb, D. R., Montgomery D. C., and W. M. Carlyle (2003). "Analysis of Supersaturated Designs". *Journal of Quality Technology*, Vol. 35, pp. 13-27.

[9] Koukouvinos, C. and K. Mylona (2008). "A Method for Analyzing Supersaturated Designs with a Block Orthogonal Structure". *Communications in Statistics - Simulation and Computation*, Vol. 37, pp. 290-300.

[10] Koukouvinos, C., Mylona K., and D. Simos (2008). "$E(s^2)$-Optimal and Minimax-Optimal Cyclic Supersaturated Designs via Multi-Objective Simulated Annealing". *Journal of Statistical Planning and Inference*, Vol. 138, pp. 1639-1646.

[11] Koukouvinos, C., Mylona K., and D. Simos (2008). "A Hybrid SAGA Algorithm for the Construction of $E(s^2)$-Optimal Cyclic Supersaturated Designs.". *Journal of Statistical Planning and Inference*, Vol. 139, pp. 478-485.

[12] Li, R. and D. K. J. Lin (2002). "Data Analysis in Supersaturated Designs". *Statist. Probab.Lett.*, Vol. 59, pp. 135-144.

[13] Li, R. and D. K. J. Lin (2003). "Analysis Methods for Supersaturated Designs: Some Comparisons". *Journal of Data Science*, Vol. 1, pp. 249-260.

[14] Lin, D. K. J. (1993). "A New Class of Supersaturated Designs". *Technometrics*, Vol. 35, pp. 28-31.

[15] Lin, D. K. J. (1995). "Generating Systematic Supersaturated Designs". *Technometrics*, Vol. 37, pp. 213-225.

[16] Liu, Y. F. and A. M. Dean (2004). "k-Circulant Supersaturated Designs". *Technometrics*, Vol. 46, pp. 32-43.

[17] Liu, M. and R. Zhang (2000). "Construction of $E(s^2)$ Optimal Supersaturated Designs Using Cyclic BIBDs". *Journal of Statistical Planning and Inference*, Vol. 91, pp. 139-150.

[18] Lu, X. and X. Wu (2004). "A Strategy of Searching Active Factors in Supersaturated Screening Experiments". *Journal of Quality Technology*, Vol. 36, pp. 392-399.

[19] Miller, A.J. (2002). *Subset Selection in Regression*, Chapman & Hall/CRC, Boca Raton.

[20] Nguyen, N. K. (1996). "An Algorithmic Approach to Constructing Supersaturated Designs". *Technometrics*, Vol. 38, pp. 69-73.

[21] Nguyen, N. K. and C. S. Cheng (2008). "New $E(s^2)$-Optimal Supersaturated Designs Constructed From Incomplete Block Designs". *Technometrics*, Vol. 50, pp. 26-31.

[22] Phoa, F.K.H., Y.-H. Pan and H. Xu (2009). "Analysis of Supersaturated Designs via Dantzing Selector". *Journal of Statistical Planning and Inference*, Vol. 139, pp. 2362-2372.

[23] Rao, C.R. and Y. Wu (2001). "On Model Selection (With Discussion)". *In: P. Lahiri (ed.), Institute of Mathematical Statistical Lecture Notes - Monograph Series*, Vol. 38, pp. 1-64.

[24] Ryan, K.J. and D.A. Bulutoglu, (2007). "$E(s^2)$-Optimal Supersaturated Designs with Good Minimax Properties. *Journal of Statistical Planning and Inference*, Vol. 137, pp. 2250-2262.

[25] Satterthwaite, F. E. (1959). "Random Balance Experimentation (With Discussions)". *Technometrics*, Vol. 1, pp. 111-137.

[26] Schwarz, G. (1978). "Estimating the Dimension of a Model". *The Annals of Statistics*, Vol. 6, pp. 461-464.

[27] Tang, B. and C. F. J. Wu (1997). "A Method for Constructing Supersaturated Designs and its $E(s^2)$-Optimality". *Canadian J. Statist.*, Vol. 25, pp. 191-201.

[28] Wang, P. C. (1995). "Comments on Lin". *Technometrics*, Vol. 37, pp. 358-359.

[29] Westfall, P. H., Young, S. S., and D. K. J. Lin (1998). "Forward Selection Error Control in the Analysis of Supersaturated Designs". *Statistica Sinica*, Vol. 8, pp. 101-117.

Letter to the Referees concerning our revision LSSP-10077 of the submiss
A Variable Selection Method for Analyzing Supersaturate
by C. Koukouvinos, K. Mylona and A. Skountzou
October, 2010

This letter describes the revision LSSP-10077 of our submission entitled A Variab Selection Method for Analyzing Supersaturated Designs after the comments of the referee and associate editor received September 07, 2010. Taking into consideration the points you kindly made, we report point to point response as follows.

Response to Associate Editor

The referee is still asking about the modification of AIC and BIC. On what basis was this modification made? You say that it worked better than other modifications, but give no evidence. It seems to me that, in g increase Type I error then Type II error will decrease. So what di choice of modification? Is there a statistical argument divided by p? The referee does not see that the new method is better than the original, careful discussion.

ð· Inspired from the work of Phoa et al. [22] we propose the and BIC with penalties of the model complexity than the original versions The use of a linear penalty leads to more provident models, modify the original versions of the criteria in order to imp application with the proposed method. Some further comments are added i text in order to clarify this choice, which is verified mainly from the simula Moreover, we rerun the simulations more carefully and the the conventional versions give low values of Type I (thus, low Type rates), while the modified versions, modAIC and modBIC perfor many cases, they outperform SCAD run into blocks.

It is not clear to me which methods are applied to the blocks and which to the entire de of the design. If the mod(AIC) is applied blockwise and SCAD is applied to th design then we are comparing different criteria as well as different meth

ð· All the methods are applied to the whole design and not to the blocks of the design. In page 6 of the previous manuscript, this point is totally clarified in Section 3 of version.

When you made the subdivisions into many blocks, as in Tabl the number of factors per block is smaller than the number of runs, so an ordinary analysis of variance could be done for a main effects model. Perhaps you should be looking at la here? Please add into the captions of the tables how many blocks of factors are analysed per block of the design.

- A larger design with 66 factors is included. In one of simulations two blocks subdivided into three blocks; thus, we have 6 blo each block of the design. The number of the total factors and the number of the active factors per block are added into the captions of the ta

The referee points out that it would be better to take the factor effects from a distribution around zero rather than setting them equal to zero.

- We rerun the simulation study generating the coefficients of active factors from N(0, 0.05).

Page 5. You talk about "up to n/2 factors" but I think you show "up to n/4" in your tables?

- The tables show the results up to n/2 factors, where n den of the design.

Page 5 end. You say "Since we used two blocks", but you have extended your work to more than two blocks.

- This phrase is corrected.

Did you use a version of SCAD obtained from previous authors or did you program it your own, did you also include the GCV step and the same parameter values as in [12] ar Please clarify.

- Further comments on the application of SCAD are added parameter values as in [9].

Page 9, line 3. It would be clearer to say "factors" instead of "runs" since this is the problem with the information matrix I think (not the number of runs). Can it be run on a very large number of variables? You do not seem to have any example with a large number of factors.

- Right, it was a typographical error; the correct phrase is the number of factors enlarges. As it is mentioned before, we present the results of a design with 66 factors in 12 runs in the case of the subdi

Response to Referee #1

First of all, what can we learn from the simulation (Tables I to VI)? Does th "fact" that the proposed method is indeed superior at all? [I don't see mu minor point is about the simulation: the inactive effects should not be assig zero, instead it should be a small random effect, such as N(0,0.5); while th shouldn't all be positive.]

- We rerun the simulation study generating the inactive effects from N(0,0.5) and selecting the coefficients of the active factors 2.0], assigning both negative and positive signs to the co method is compared to SCAD block. We believe the new simulation results reveal the effectiveness of the proposed method, since i two modified criteria behave better than the conventional they compete block SCAD. Especially, when the proposed modAIC, it gives better results than SCAD in most of the cases.

Second, The proposed criteria, modAIC and modBIC, need explanation motivation for why so or even their general properties. The authors indicated that (top of p.4) "Several simulations showed that..." This is not convincing at all. Why this is a good way to "standardize", by the way.

- Motivated from the work of Phoa [22] we propose the two modific AIC and BIC with different penalties of the model comple versions. Some further comments are added in the text i choice, which is verified mainly from the simulation study. The use of linear penalty leads to overfitted models, so we tried to modify the origin of the criteria in order to improve the results of their application method.